Mathematical Geology, Vol. 27, No. 5, 1995

Infill Sampling Criteria to Locate Extremes[1]

Alan G. Watson[2] and Randal J. Barnes[3]

Three problem-dependent meanings for engineering "extremes" are motivated, established, and translated into formal geostatistical (model-based) criteria for designing infill sample networks. (1) Locate an area within the domain of interest where a specified threshold is exceeded, if such areas exist. (2) Locate the maximum value in the domain of interest. (3) Minimize the chance of areas where values are significantly different from predicted values. An example application on a simulated dataset demonstrates how such purposive design criteria might affect practice.

KEY WORDS: geostatistics, sample design.

INTRODUCTION

Consider a road cut along which a new highway is planned. When the cut was considered originally, soil and rock samples were taken at irregular intervals along the alignment; sample locations were determined subjectively by an onsite specialist. The results of these initial samples indicated that there might be problems in the construction of the road owing to a few especially weak zones; but, it is not clear just how pervasive these zones are. The road contractors require a more intensive survey before they are prepared to bid for the construction job, as those portions of the highway which traverse the weak materials require special treatment at considerably greater cost than the rest of the alignment.

The problem faced by the highway authority is one of infill sampling to characterize extreme values. In this particular application, an extreme value is defined by a geomechanical threshold value, but an engineer's use of the term "extreme" implies different things depending on the application. The objective of this paper is to establish some problem-dependent meanings for "extreme," and to translate these meanings into formal criteria for designing infill obser-

[1] Received 18 July 1994; revised 14 March 1995.

[2] CSIR, Informationtek Division, Pretoria, South Africa.

[3] Department of Civil Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

vation networks, recognizing the inherent spatial correlation of geologic data (i.e., measurements taken close together tend to be more similar than measurements taken farther apart).

This paper addresses the following general problem. Based upon a set of existing observations in and around the area of interest, and a statistical model describing the spatial variation of these observations, select a set of new sample locations to achieve a stated purpose. This problem is known by the clumsy title: *model-based infill sample network design*. For brevity, this problem title will be shortened to *infill design* in this paper.

BACKGROUND

With a formalized theoretical basis and a record of relative success, geostatistics has been embraced by many fields. The geostatistical formulation of infill design, which is applied in this paper, provides a common vocabulary, a framework for stochastically modeling the unknown values between the existing observations, and a mathematically consistent mechanism for incorporating new information in the model as it becomes available.

During the past 25 years a significant quantity of work has been published on infill design using a geostatistical framework. For example, Davis and Dvoranchik (1971), Duckstein and Kisiel (1971), Rodríguez-Iturbe and Mejia (1974), Bras and Rodríguez-Iturbe (1976a, 1976b), Bras and Colon (1978), Attanasi and Karlinger (1979), Davis, Duckstein, and Krysztofowicz (1979), Dawdy (1979), Gershon (1983), Rouhani (1985), Barnes (1989), and Cressie (1991, section 5.6) all discuss sample network design using a geostatistical framework. More recently, Thompson (1992) devotes an entire book to spatial sampling and estimation, in which the geostatistical framework is important.

To the exclusion of almost all other criteria, these recent scholarly publications on infill design have concentrated on minimizing the estimation variance of the areal mean. Nonetheless, in geological engineering settings (and many others as well), estimation of the areal mean is rarely the objective of infill sampling. Geological engineering designs usually are governed by the extraordinary, rather than mean values: for example, extreme values and extreme deviation from expectations.

In the following three sections, three different meanings for ''extreme'' are identified and motivated. In each situation a purposive sampling strategy is constructed, and the infill design implications of these definitions are investigated. The consequences of various simplifications (e.g., distributional assumptions) also are considered. Because the problem at hand is infill sampling random and random stratified sampling are not appropriate.

NOTATION

The following notation will be used consistently throughout this presentation. Whenever a stochastic model is implied (e.g., an expected value is computed), the geostatistical model as defined by Cressie (1991, section 2.1) is adopted.

A is the area of interest.

$z(x)$ is the observed value at location $\mathbf{x}$.

S is the set of n existing samples, $\{z(\mathbf{y}_1), \ldots, z(\mathbf{y}_n)\}$, taken at locations $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ in and around A.

$Z(\mathbf{x})$ is the random variable representing the unobserved value at location $\mathbf{x}$.

$\hat{z}(\mathbf{x})$ is the modeled (i.e., predicted) value at location $\mathbf{x}$.

F( ) is the cumulative distribution function of the identified random variable or variables; usually the specified distribution function will be multivariate and conditional, as indicated by the index and arguments.

Pr[ ] is the probability of the identified event.

E[ ] is the expected value of the identified event.

Var[ ] is the variance of the identified event.

Using this notation, infill design can be described as follows. The design starts with a predefined area of interest, A, and a set of n existing observations, $S \equiv \{z(\mathbf{y}_1), \ldots, z(\mathbf{y}_n)\}$, taken at locations $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ in and around A. The objective is to determine an appropriate configuration of locations, $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, at which to take m new observations. These m new observation may be taken as a single batch of size m or as a sequential progression of smaller batches, even individuals.

LOCATING THRESHOLD-BOUNDED EXTREMES

Consider the following geological engineering problem, and the qualitative engineering objective that it suggests.

Example Problem. A potentially contaminated site has been identified as a possible health hazard as a result of purported toxic contamination. Legal recourse is available if a regulatory health and safety threshold is exceeded on the property.

Engineering Objective. Locate an area within the domain of interest where a specified threshold is exceeded, if such an area exists.

This example problem and the associated engineering objective suggest the first problem-dependent meaning for "extreme."

Definition. A *threshold-bounded extreme* is a value, $z(\mathbf{x})$, that exceeds a specified threshold value, T.

The threshold, T, typically is a function of external considerations (e.g., maximum safe concentrations of contaminants as defined by regulation or minimum economic ore grades). The problem is *not* one of locating the contours surrounding areas which exceed the threshold, a question that was considered by Veneziano and Kitanidis (1982) and Aspie and Barnes (1990). Rather, the question is: ''Is the threshold exceeded anywhere within the area of interest?'' Implicit in this question is the assumption that none of the existing observations exceeds the threshold, for otherwise the question already is answered in the affirmative and the sampling objective would switch from identification to characterization.

An appropriate, nonzero, sample support also is implicit in this definition. If the physical sample support were allowed to shrink to an arbitrary infinitesimal size, any concentration is possible as long as there is one molecule of the target compound present. The following discussion presumes that the support for the collected samples is appropriate and commensurate with the support used in defining the specified threshold.

Using the geostatistical model, the sampling objective can be posed as the following stochastic optimization problem.

Sampling Objective. Given the set of n existing observations, $S \equiv \{z(\mathbf{y}_1, \ldots, z(\mathbf{y}_n)\}$, take a set of m new observations, $\{Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_m)\}$, at those locations within the area of interest, $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, which maximize the probability that at least one of the new observations exceeds the specified threshold value, T.

This objective may be written formally as:

$$\max_{\mathbf{x}_1 \ldots \mathbf{x}_m \in A} \{\mathbf{Pr}[\text{At least one } Z(\mathbf{x}_i) > T \mid S]\}$$

$$= \max_{\mathbf{x}_1 \ldots \mathbf{x}_m \in A} \{1 - \mathbf{Pr}[\text{All } Z(\mathbf{x}_i) \leq T \mid S]\}$$

$$= \max_{\mathbf{x}_1 \ldots \mathbf{x}_m \in A} \{1 - \mathbf{F}_{Z(\mathbf{x}_1) \ldots Z(\mathbf{x}_m) \mid S}(T, \ldots, T)\}$$

which is equivalent to

$$\min_{\mathbf{x}_1 \ldots \mathbf{x}_m \in A} \{\mathbf{F}_{Z(\mathbf{x}_1) \ldots Z(\mathbf{x}_m) \mid S}(T, \ldots, T)\} \tag{1}$$

where $\mathbf{F}(\ )$ is the appropriate multivariate conditional cumulative distribution function.

Equation (1) presents a general mathematical objective for locating threshold-bounded extremes. It captures the essence of the sampling goal, is applicable for any number new samples, and is correct for any multivariate distribution.

Simplifying the Model

To exercise the objective and investigate the resulting sampling strategy in a simple setting consider the case where the batch size is one (i.e. $m = 1$). Then the sampling objective [Eq. (1)] may be restated as:

$$\min_{\mathbf{x} \in A} \{\mathbf{F}_{Z(\mathbf{x})|S}(T)\}$$

If, in addition, a homogeneous Gaussian random field model is embraced, then the marginal conditional distribution may be parameterized in terms of the conditional mean and the conditional variance:

$$\mu_{Z(\mathbf{x})|S} = \mathbf{E}[Z(\mathbf{x})|S]$$

$$\sigma^2_{Z(\mathbf{x})|S} = \text{Var}\ [Z(\mathbf{x})|S]$$

A normalized "z-score" then may be defined for the specified threshold as

$$\zeta(\mathbf{x}) = \frac{T - \mu_{Z(\mathbf{x})|S}}{\sigma_{Z(\mathbf{x})|S}}$$

Because the standard Gaussian cumulative distribution function is a strictly increasing function of its argument, it suffices to take a new observation, $\{Z(\mathbf{x})\}$, at that location, $\mathbf{x}$, which minimizes the z-score. Thus, the objective can be written

$$\min_{\mathbf{x} \in A} \{\zeta(\mathbf{x})\} \tag{2}$$

which is a remarkably simple formula, amenable to efficient application, quickly explained, and easily interpreted.

Observations

(1) Although, for the purposes of this discussion, a threshold-bounded extreme was defined to be a value that exceeds a specified upper threshold, it could well have been defined as a value that falls short of a specified lower threshold.

(2) The popular nonparametric (indicator) kriging methods (e.g., Journel, 1983), and the distribution-free simulation based methods (e.g., Journel and Alabert 1988), cannot estimate the necessary conditional multivariate probabilities [i.e., those required in Eq. (1)], because the threshold is greater than the largest sample value.

(3) A multivariate distributional model is necessary for the implementation of this objective. Furthermore, it is necessary to embrace a distributional model that can not be proven correct by the existing observations.

The multigaussian model (e.g., Verly, 1983) offers a simple, pragmatic heuristic method for assessing the necessary conditional multivariate probabilities. It is possible to develop a distributional model using one of the various forms of disjunctive kriging (e.g., Armstrong and Matheron, 1986a, 1986b).

(4) When using the simple, one-sample, Gaussian model [i.e., Eq. (2)], it can be seen that the size of the threshold, T, determines whether infill sampling concentrates around existing large values, or around holes in the sampling. The objective balances between the chance of locating an extreme value by filling in the holes (locations with larger conditional variances), and the chance of locating an extreme value near the current known larger values (locations with larger conditional means).

(5) Equation (2) suggests a "quick-and-dirty" implementation of the threshold-bounded extreme objective. Apply a transformation to the available data so that the resulting univariate histogram is approximately Gaussian. Use the ordinary kriging prediction and the ordinary kriging variance of the transformed data as estimates of the local conditional mean and variance. Apply Equation (2) and select the location with the minimum z-score. (Remember to transform the threshold, T, before computing the normalized z-scores.)

LOCATING THE REGIONAL EXTREME

Consider the following geological engineering problem, and the qualitative sampling objective that it suggests.

Example Problem. Suppose that hazardous concentrations of toxic contamination have been identified in a suburban district. The contaminant poses an immediate threat to the health of the residents, so it is important to prioritize the cleanup, starting with those areas where the concentration is highest.

Engineering Objective. Locate the maximum concentrations in the district.

This example and the associated engineering objective suggest the second problem-dependent meaning for "extreme."

Definition. The *location of the regional extreme* is that subset of the area of interest where the highest value is achieved.

Observe that this definition of the regional extreme permits the existence of multiple points, not necessarily contiguous, at which the realization may attain its maximum value. Except in special circumstances (e.g., an ordinal random process model), however, the regional maximum is attained at a single point. Consequently, the probability of placing a new observation at precisely that location is zero. Nonetheless, the sense of the sampling objective, which can be achieved by taking point observations, is the maximization of the sample maximum order statistic.

Sampling Objective. Given the set of n existing observations, $S \equiv \{z(\mathbf{y}_1), \ldots, z(\mathbf{y}_n)\}$, take a set of m new observations, $\{Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_m)\}$, at those locations, $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, which maximize the expected value of the largest observation after the new observations have been taken.

This objective may be written formally as

$$\max_{\mathbf{x}_1 \ldots \mathbf{x}_m \in A} \{\mathrm{E}[Z_{(n+m)} | S]\} \tag{3}$$

where $Z_{(n+m)}$ is the maximum sample value out of the n existing samples and the m new samples; although the first n observations are known, $Z_{(n+m)}$ is as a random variable.

Equation (3) presents a general mathematical objective for locating the regional extreme. It captures the essence of the sampling goal, is applicable for any number of new samples, and is correct for any multivariate distribution.

Simplifying the Model

To exercise the objective and investigate the resulting sampling strategy in a simple setting, consider the case where the batch size is one (i.e., $m = 1$). The objective function [Eq. (3)] then may be restated as:

$$\max_{\mathbf{x} \in A} \{\mathrm{E}[Z_{(n+1)} | S]\} = \mathrm{E}[\max_{\mathbf{x} \in A} [Z(\mathbf{x}), z_{(n)}] | S] \tag{4}$$

where $z_{(n)}$ is the maximum existing sample value; $z_{(n)}$ is not a random variable. If, in addition, the generating random field is gaussian and homogeneous, then the marginal conditional distribution may be parameterized in terms of the conditional mean and the conditional variance:

$$\mu_{Z(\mathbf{x})|S} = \mathrm{E}[Z(\mathbf{x}) | S]$$

$$\sigma^2_{Z(\mathbf{x})|S} = \mathrm{Var}\ [Z(\mathbf{x}) | S]$$

A local normalized "z-score" for the current largest sample value may be defined as:

$$\xi(\mathbf{x}) = \frac{Z_{(n)} - \mu_{Z(\mathbf{x})|S}}{\sigma_{Z(\mathbf{x})|S}}$$

Explicit incorporation of the standard normal density function into Equation (4), followed by integration and algebraic rearrangement yields a concise statement of the single-sample, Gaussian-based, objective function:

$$\mathrm{E}[Z_{(n+1)} | S] = \mu_{Z(\mathbf{x})|S} + W(\xi(\mathbf{x})) \cdot \sigma_{Z(\mathbf{x})|S}$$

where $W(\ )$ is the peculiar function

$$W(\xi(\mathbf{x})) = \phi(\xi(\mathbf{x})) + \xi(\mathbf{x}) \cdot \Phi(\xi(\mathbf{x}))$$

$\phi(\xi)$ represents the standard Gaussian density function, and $\Phi(\xi)$ represents the standard Gaussian cumulative distribution function (see the Appendix for a derivation). Thus, the objective can be written:

$$\max_{\mathbf{x} \in A} \{\mu_{Z(\mathbf{x})|S} + W(\xi(\mathbf{x})) \cdot \sigma_{Z(\mathbf{x})|S}\} \tag{5}$$

which is a remarkably simple formula, amenable to efficient application, and easily interpreted.

Observations

(1) Although, for the purposes of this discussion, a regional extreme was defined to be a regional maximum, it equally well could have been defined as a regional minimum.

(2) As in the situation of the threshold-bounded extreme, the popular nonparametric (indicator) kriging methods (e.g., Journel, 1983), and the distribution-free simulation based methods (e.g., Journel and Alabert 1988), can not estimate the necessary conditional expectation [(i.e., Eq. (3)].

(3) A multivariate distributional model is necessary for the implementation of this objective. Again, the multigaussian model (e.g., Verly, 1983) offers a simple, pragmatic heuristic method for assessing the necessary probabilities [i.e., Eq. (3)]. A disjunctive kriging model could be applied.

(4) When using the simple, one-sample, Gaussian model, the function, $W(\xi(\mathbf{x}))$, provides a mechanism for balancing between sampling near the existing sample maximum (locations with larger conditional means), and sampling in holes in the current sample network (locations with larger conditional variances). When $\xi(\mathbf{x})$ is large $W(\xi(\mathbf{x}))$ is large and almost linear in $\xi(\mathbf{x})$. As $\xi(\mathbf{x})$ becomes small $W(\xi(\mathbf{x}))$ rapidly tends to zero. Consequently, when the area of interest is well covered and the existing sample maximum, $z_{(n)}$, is larger than the other observations, the new observation will be taken near to $z_{(n)}$; but, when no one observation is larger than the others and the sample network has obvious holes, the new observation will be placed in one of those holes.

(5) Equation (5) suggests a "quick-and-dirty," implementation of the objective. Apply a transformation to the available data so that the resulting univariate histogram is approximately Gaussian. Use the ordinary kriging prediction and the ordinary kriging variance of the transformed data

as estimates of the local conditional mean and variance. Select the location that maximizes Equation (5).

MINIMIZING SURPRISES

Consider the following geological engineering problem, and the qualitative sampling objective that it suggests:

Example Problem. Although the average ore grade in a mine is an important measure of the economic viability of the mine, the local predictability of the ore grade may be as important in the generation of correct mine plans.

Engineering Objective. Minimize the chance of encountering a region of ore whose grade is unexpectedly different from its predicted value.

What the miners are concerned with here are "surprises": locations where reality differs radically from the predictive model built to describe it. This example problem and associated engineering objective suggest the third problem-dependent meaning for "extreme" (in this situation extreme discrepancies are of concern, not extreme values).

Definition. A *surprise* is a true value that deviates significantly from its predicted value.

"Significant deviation" is subjective and problem-dependent, yet it determines how much and where additional sampling is required. A formulation that captures much of the essence of this objective follows.

Sampling Objective. Given the set of n existing observations, $S \equiv \{z(\mathbf{y}_1), \ldots, z(\mathbf{y}_n)\}$, take a set of m new observations, $\{Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_m)\}$, at those locations, $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, which minimize the maximum probability that a true value deviates significantly from its predicted value; where the predicted value is based upon the existing samples and the new observations.

This objective may be written formally as

$$\min_{\mathbf{x}_1 \ldots \mathbf{x}_m \in A} \; \max_{\mathbf{v} \in A} \{\mathbf{Pr}[|Z(\mathbf{v}) - \hat{z}(\mathbf{v})| > t | S]\} \tag{6}$$

where $\mathbf{v}$ is a generic location within the area of interest, not necessarily a sample location, $Z(\mathbf{v})$ is the random variable representing the unobserved value at location $\mathbf{v}$, $\hat{z}(\mathbf{v})$ is the modeled (i.e., predicted) value at location $\mathbf{v}$, and t is a specified tolerance level defining "significant deviation."

Evaluation of such an objective function requires a multivariate conditional distributional model and a quantification of "significant deviation" in terms of a tolerance, t. Furthermore, as this objective purports to minimize surprises (in some sense), the conditional distribution should include the uncertainty in selecting the geological, geochemical, and probabilistic models. Many surprises are the result of poor physical and probabilistic models: failure to recognize important subpopulations, inappropriate anisotropy in the variogram models, spatial patterns that are not captured by second-moment statistics, etc.

Equation (6) presents a general mathematical objective for minimizing surprises. It is applicable for any number new samples, and is correct for any multivariate distribution.

Simplifying the Model

When the predicted value is the conditional expectation, Chebychev's inequality gives an upper bound on the required probability [i.e., Eq. (6)] as an increasing function of the conditional variance, (e.g., Manoukian, 1986, p. 11). This suggests the use of the conditional variance as an approximate distribution-free surrogate for the probability:

$$\min_{\mathbf{x}_1 \ldots \mathbf{x}_m \in A} \ \max_{\mathbf{v} \in A} \{\mathrm{Var}\,[Z(\mathbf{v}) | S \text{ and } \mathbf{x}_1 \ldots \mathbf{x}_m]\} \tag{7}$$

This simplified objective function for a candidate set of observation locations, $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, is computed as the maximum conditional variance over the area of interest assuming that locations $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ already have been observed. The set of locations that results in the smallest expected posterior uncertainty (quantified by the maximum conditional variance) is selected as the batch of new locations. The goal is to select that set of locations that will minimize the worst chance of a subsequent surprise; this is a mini-max decision rule.

Observations

(1) When using a multivariate Gaussian distribution model the objective function depends upon only the observation *locations* and the underlying *covariance* structure. Thus, in this situation, the observed values do not enter into the evaluation (e.g., Journel and Huijbregts, 1978, p. 308).

(2) Minimization of the maximum kriging variance was proposed by Burgess, Webster, and McBratney (1981). The emphasis of their paper however is not on infill sampling, but on the design of new sampling programs: specifically, they were interested in the effect which observation spacing has on the maximum variance for regular sampling patterns.

(3) Suppose the underlying covariance structure is such that there exists a separation distance, *r*, beyond which any two points are uncorrelated (an example of such a correlation structure is a spherical variogram with range *r*). Then, if there exist $m + 1$ or more places in the area of interest which are not "covered" by the existing observations (i.e., there are no measurements within a distance *r*), and which cannot be covered simultaneously by a batch of *m* new observations, the ob-

jective function is flat. In practice, thus the area of interest must be covered fully before any attempt is made to minimize surprises using this objective function.

(4) Equation (7) suggests a "quick-and-dirty," implementation of this objective. Apply a transformation to the available data so that the resulting univariate histogram is approximately gaussian. Use the ordinary kriging prediction and the ordinary kriging variance of the transformed data as estimates of the local conditional mean and variance. Select the location that minimizes the maximum posterior kriging variance of the transformed field. Although this approximation, in fact, may be crude for data that can not be modeled adequately as multivariate Gaussian random fields, the approximation offers an easily computed heuristic objective capturing the essential geometric details of the sampling pattern. Nonetheless, this "quick-and-dirty" approach is less attractive than those given for the previous objectives, because the resulting objective function does not incorporate information about the local conditional mean. Furthermore, the "significant deviation" used in identifying surprises is in terms of the transformed data—which may be difficult to interpret, or simply inappropriate.

AN EXAMPLE

Each of the three sampling objectives discussed here was applied to a set of simulated data. These examples demonstrate that the choice of an appropriate infill design objective can be significant. The resulting sample configurations consistently are different from one another in visibly obvious ways.

The computer experiments were performed on simulated data generated with the following parameters:

• Distribution:	multivariate lognormal
• Log mean:	0.0
• Log semi-, variogram:	$\gamma(h) = \begin{cases} 1.2 \cdot \left[1.5\left(\dfrac{h}{12.5}\right) - 0.5\left(\dfrac{h}{12.5}\right)^3\right] & \text{if } h < 12.5 \\ 1.2 & \text{if } h \geq 12.5 \end{cases}$
• Grid spacing:	1.0 × 1.0
• Grid size:	25 × 25

The simulation was accomplished using the well-known method employing the decomposition of the variance/covariance matrix (e.g., Naylor and others, 1966,

p. 97–101). Analysis of the exhaustive data (625 values) yields the following statistics:

• Minimum:	0.073
• Median:	1.024
• Maximum:	11.74
• Arithmetic average:	1.643
• Standard deviation:	1.875

Figure 1 shows a contour plot of the exhaustive data: 625 observations on a 25 × 25 grid. Figure 2 shows the equivalent contour plot of the log-transformed exhaustive data.

For each of the three sampling objectives, the demand was for an additional five observation locations given a set of 20 existing observations. The locations and values of the existing 20 observations are shown in Figure 3. The new observations were selected as the best batch of five from the 605 unsampled grid nodes. "Best" was defined by Equations (1) and (3), not the "quick-and-

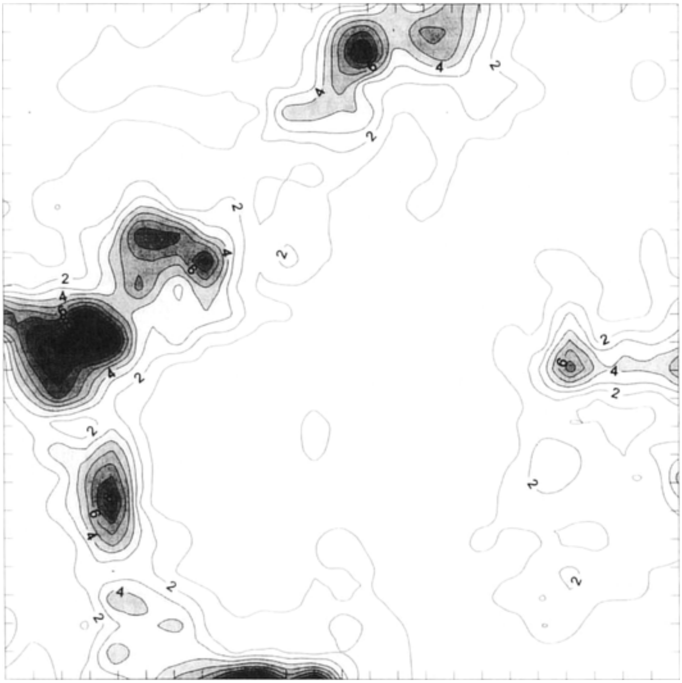

Figure 1. Contours of exhaustive dataset: 625 observations on 25 × 25 grid.

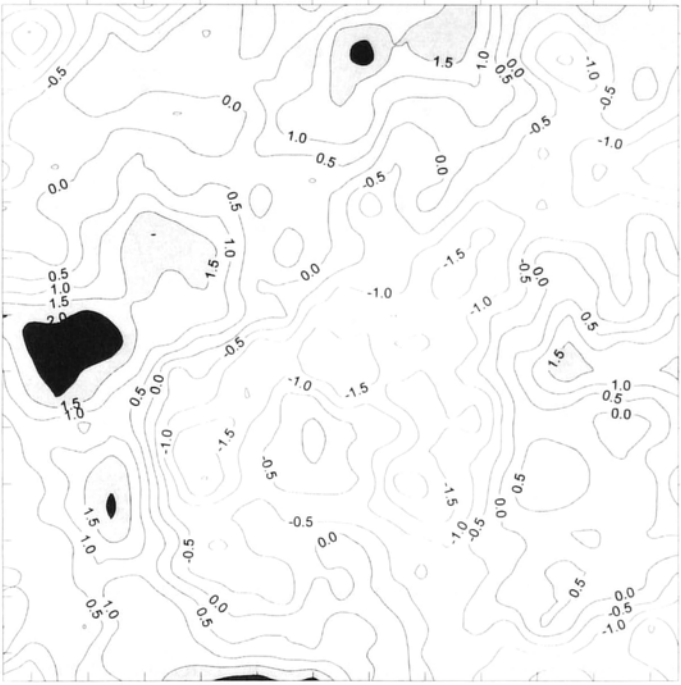

Figure 2. Contours of log-transformed exhaustive data set: 625 observations on 25 × 25 grid.

dirty'' surrogates, and Equation (6). The computational effort involved in such a search was reduced significantly using the techniques presented in Barnes and Watson (1992).

This experiment was carried out in a congenial multigaussian setting, with all of the model-based assumptions known to be appropriate a priori. Because the data were simulated as a realization of a multivariate lognormal distribution, the three infill sampling objectives were applied to the log-transformed observations. The spatial covariance function used in evaluating the various objectives was computed from the prespecified variogram and not estimated using the 20 observations.

Locating Threshold Bounded Extremes

In this example, the operational objective is to maximize the probability of observing a value exceeding a threshold [Eq. (1)]. A value of 4.2 was selected

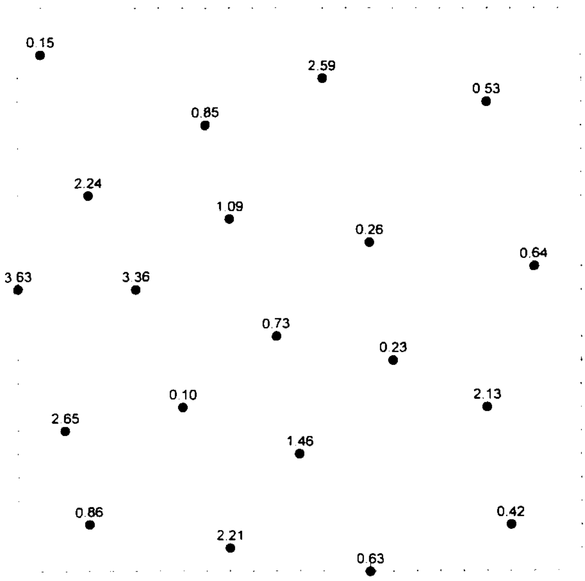

Figure 3. Locations and values of 20 existing observations.

for the threshold (approximately equal to the 95 percentile of the exhaustive data). The results are shown in Figure 4.

To interpret the observation placement for this objective physically, consider an estimated value grid overlaid with an estimation variance grid. Then, new observations will tend to be placed in two sorts of locations: (1) where the estimated value is large and the estimation variance is moderate; and, (2) where the estimated value is moderate, but the estimation variance is large.

Locating the Regional Extreme

In this example, the operational objective is to maximize the expected posterior sample maximum [Eq. (3)]. The results are shown in Figure 5.

New observations are placed in the middle of the west edge (this is in the neighborhood of the maximum existing sample), and in the middle of the top edge (this is an area having few samples, but with the potential to be large).

Minimizing Surprises

In this example, the operational objective is to minimize the maximum local posterior estimation variance [Eq. (7)]. The results are shown in Figure 6.

New observations tend to be placed in holes in the existing sample network.

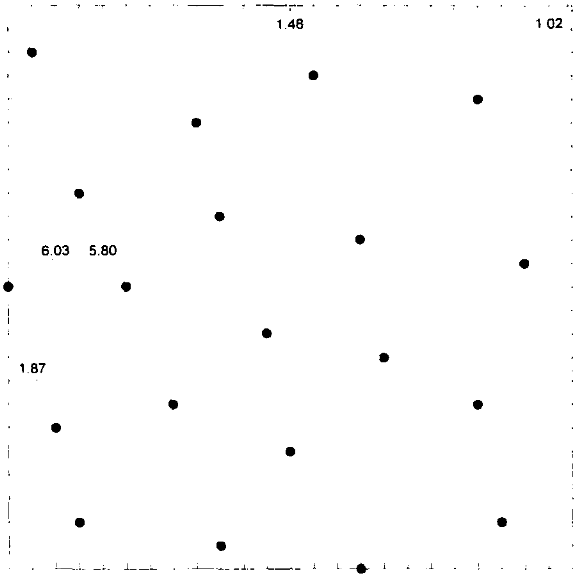

Figure 4. Locations and values of five new observations placed using threshold bounded extreme objective: $T = 42$. Locations of existing 20 observations also are shown.

The locations are selected independently of the observed data, ignoring anomalies which have been detected. Not surprisingly, the new observations are located on the border and in holes in the existing sampling pattern.

Comparing the Objectives

These examples show that different objectives should lead to different sampling patterns. For the simulated (multivariate lognormal) grid, all three of the proposed objectives performed in heuristically satisfactory ways.

A quantitative comparison of the objective function values is shown in Table 1. The largest observation using the regional extreme objective significantly exceeded the largest observation using the other two objectives. Furthermore, despite the apparent similarity between the threshold bounded extreme objective and the regional maximum objective, the resulting infill designs are significantly different.

WHY NOT CONDITIONAL SIMULATION?

Conditional simulation is a powerful tool for generating derived distributions, especially when the random variable of interest is related to the random

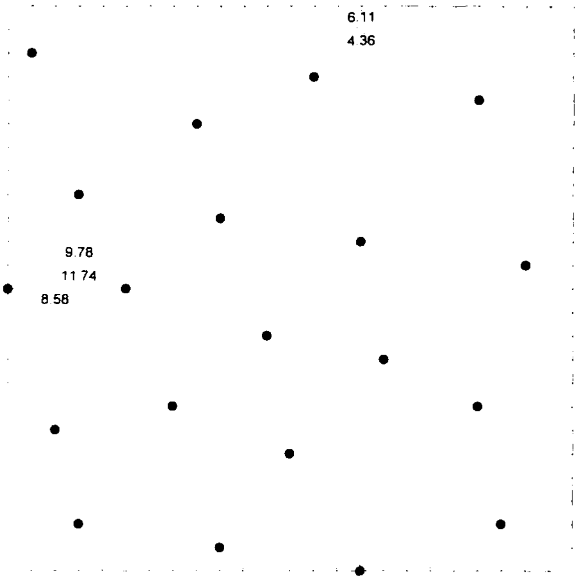

Figure 5. Locations and values of five new observations placed using regional extreme objective. Locations of existing 20 observations also are shown.

field via a complex nonlinear relationship (e.g., Deutsch and Journel, 1992). With the currently available software, conditional simulations are being applied in many fields. Why not then simply perform a series of conditional simulations and process the results with the appropriate objective to determine the additional sample locations? For example, using conditional simulation the minimizing surprises objective [Eq. (6)] could be considered explicitly instead of its simplification [Eq. (7)].

Conditional simulation offers a conceptually appealing approach for solving various infill design problems. In fact, conditional simulation can be used to evaluate the three objectives discussed in this paper. Unfortunately, such an approach is intractable computationally for many problems. Consider, it may take tens to hundreds of conditional simulations, if not more, to estimate adequately the necessary derived distribution statistics. In the simple example presented in the preceding section, the problem was to determine the best 5 locations out of 620 available candidate locations. There are 751,199,765,624 candidate combinations. Even if heuristic combinatorial search techniques are used, thousands of candidate combinations would have to be considered—and each combination could require tens to hundreds of conditional simulations during the

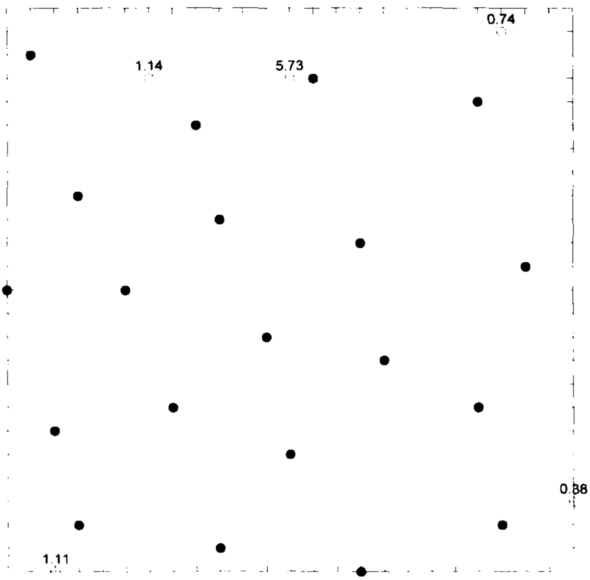

Figure 6. Locations and values of five new observations placed using minimizing surprises objective (i.e., minimizing maximum local conditional variance). Locations of existing 20 observations also are shown.

evaluation of the selected objective. With the current computer technology, and understand of conditional simulation, such a process could not be completed in a reasonable length of time.

CONCLUSIONS

Different problems and different questions engender different objectives. Using the geostatistical framework, it is possible to formulate and implement

Table 1. Quantitative Comparison of Three Sampling Objectives

	Maximum posterior observation	Maximum local posterior estimation variance of the log-transformed field
Threshold bounded extreme	6.04	0.811
Regional extreme	11.74	1.091
Minimizing surprises	5.73	0.788

various infill design strategies to characterize "extremes." The particular objective used should be selected to fit best the circumstances: it can make a significant difference.

When problem-specific objective functions can be formulated for an infill sampling application, purposive, model-based, sample network design can be a useful policy.

ACKNOWLEDGMENTS

The initial development of the ideas presented in this paper were funded, in part, by a grant from the United States National Science Foundation Grant No. NSF/MSM-8710568. The comments and constructive criticism of the referees significantly improved the readability of this paper.

REFERENCES

Armstrong, M., and Matheron, G., 1986a, Disjunctive kriging revisited: part I: Math. Geology, v. 18, no. 8, p. 711-728.

Armstrong, M., and Matheron, G., 1986b, Disjunctive kriging revisited: part II: Math. Geology, v. 18, no. 8, p. 729-742.

Aspie, D., and Barnes, R. J., 1990, Infill-sampling design and the cost of classification errors: Math. Geology, v. 22, no. 8, p. 915-932.

Attanasi, E. D., and Karlinger, M. R., 1979, Worth of data and natural disaster insurance: Water Resources Research, v. 15, no. 6, p. 1763-1766.

Barnes, R. J., 1989, Sample design for geologic site characterization, *in* Armstrong, M., ed., Geostatistics, Vol. 2: Kluwer, Dordrecht, p. 809-822.

Barnes, R. J., and Watson, A. G., 1992, Efficient updating of kriging estimates and variances: Math. Geology, v. 24, no. 1, p. 129-134.

Bras, R. L., and Colon, R., 1978, Time-averaged areal mean of precipitation: estimation and network design: Water Resources Research, v. 14, no. 5, p. 878-888.

Bras, R. L., and Rodríguez-Iturbe, I., 1976a, Network design for the estimation of areal mean of rainfall events: Water Resources Research, v. 12, no. 6, p. 1185-1195.

Bras, R. L., and Rodríguez-Iturbe, I., 1976b, Rainfall network design for runoff prediction: Water Resources Research, v. 12, no. 6, p. 1197-1208.

Burgess, T. M., Webster, R., and McBratney, A. B., 1981, Optimal interpolation and isarithmic mapping of soil properties: IV. Sampling Strategy: Jour. Soil Science, v. 32, no. 4.

Cressie, N. A. C., 1991, Statistics for spatial data: John Wiley & Sons, New York, 900 p.

Davis, D. R., Duckstein, L., and Krysztofowicz, R., 1979, The worth of hydrologic data for nonoptimal decision making: Water Resources Research, v. 15, no. 6, p. 1733-1742.

Davis, D. R., and Dvoranchik, W. M., 1971, Evaluation of the worth of additional data: Water Resources Bull., v. 7, no. 4, p. 700-707.

Dawdy, D. R., 1979, The worth of hydrologic data: Water Resources Research, v. 15, no. 6, p. 1726-1732.

Deutsch, C. V., and Journel, A. G., 1992, GSLIB: Geostatistical Software Library and user's guide: Oxford Univ. Press, New York, 340 p.

Duckstein, L., and Kisiel, C. C., 1971, Efficiency of hydrologic data collection systems: role of Type I and Type II Errors: Water Resources Bull., v. 7, no. 3, p. 592–604.

Gershon, M., 1983, Optimal drillhole location using geostatistics: Soc. Mining Engineers preprint 83-63, Littleton, Colorado, unpaginated.

Journel, A. G., 1983, Nonparametric estimation of spatial distributions: Math. Geology, v. 15, no. 3, p. 445–468.

Journel, A. G., and Alabert, F., 1988, Non-gaussian data expansion in the earth sciences: Terra Review, v. 1, no. 2, p. 123–134.

Journel, A. G., and Huijbregts, C., 1978, Mining geostatistics: Academic Press, London, 600 p.

Manoukian, E. B., 1986, Modern concepts and theorems of mathematical statistics: Springer-Verlag, New York, 156 p.

Naylor, T. H., Baintfy, J. L., Burdick, D. S., and Chu, K., 1966, Computer simulation techniques: John Wiley & Sons, New York, 352 p.

Rodríguez-Iturbe, I., and Mejia, J. M., 1974, The design of rainfall networks in time and space: Water Resources Research, v. 10, no. 4, p. 1185–1195.

Rouhani, S., 1985, Variance reduction analysis: Water Resources Research, v. 21, no. 6, p. 837–846.

Thompson, S. K., 1992, Sampling: John Wiley & Sons, Inc., New York, 343 p.

Veneziano, D., and Kitanidis, P. K., 1982, Sequential sampling to contour an uncertain function: Math. Geology, v. 14, no. 5, p. 387–404.

Verly, G., 1983, The multigaussian approach and it applications to the estimation of local reserves: Math. Geology, v. 15, no. 3, p. 263-290.

APPENDIX

Given a known constant β and a continuous random variable Z, with mean μ_z, density function $f_z(z)$, and cumulative distribution function $F_z(z)$, we can compute

$$\mathrm{E}[\max(Z, \beta)]$$

$$= \int_{-\infty}^{\infty} \max(\xi, \beta) f_z(\xi)\, d\xi$$

$$= \int_{\beta}^{\infty} z f_z(t)\, dt + \int_{-\infty}^{\beta} \beta f_z(t)\, dt \tag{A1}$$

$$= \int_{-\infty}^{\infty} t f_z(t)\, dt - \int_{-\infty}^{\beta} t f_z(t)\, dt + \int_{-\infty}^{\beta} \beta f_z(t)\, dt$$

$$= \mu_z - \int_{-\infty}^{\beta} t f_z(t)\, dt + \beta \mathbf{F}_z(\beta)$$

If random variable Z is Gaussian, with standard deviation σ_z, then Equation (A1) can be simplified further to

$$\begin{aligned}
&\mathbf{E}[\max(Z, \beta)] \\
&\quad = \mu_z - \int_{-\infty}^{\beta} tf_z(t)\, dt + \int_{-\infty}^{\beta} \beta f_z(t)\, dt \\
&\quad = \mu_z - \int_{-\infty}^{\beta} (t - \mu_z)f_z(t)\, dt + \int_{-\infty}^{\beta} (\beta - \mu_z)f_z(t)\, dt \qquad \text{(A2)} \\
&\quad = \mu_z + \left[\int_{-\infty}^{\beta} \frac{(\beta - \mu_z)}{\sigma_z} f_z(t)\, dt - \int_{-\infty}^{\beta} \frac{(t - \mu_z)}{\sigma_z} f_z(t)\, dt \right] \sigma_z \\
&\quad = \mu_z + [\xi\Phi(\xi) + \phi(\xi)]\sigma_z
\end{aligned}$$

where ξ is the normalized "z-score" of β

$$\xi = \frac{\beta - \mu_z}{\sigma_z}$$

$\Phi(\xi)$ is the standard normal cumulative distribution function, and $\phi(\xi)$ is the standard normal density function. The last step in Equation (A2) follows from a peculiar behavior of the standard normal density function:

$$\int_{-\infty}^{\beta} t\phi(t)\, dt = -\phi(\beta)$$

which can be verified by integration.